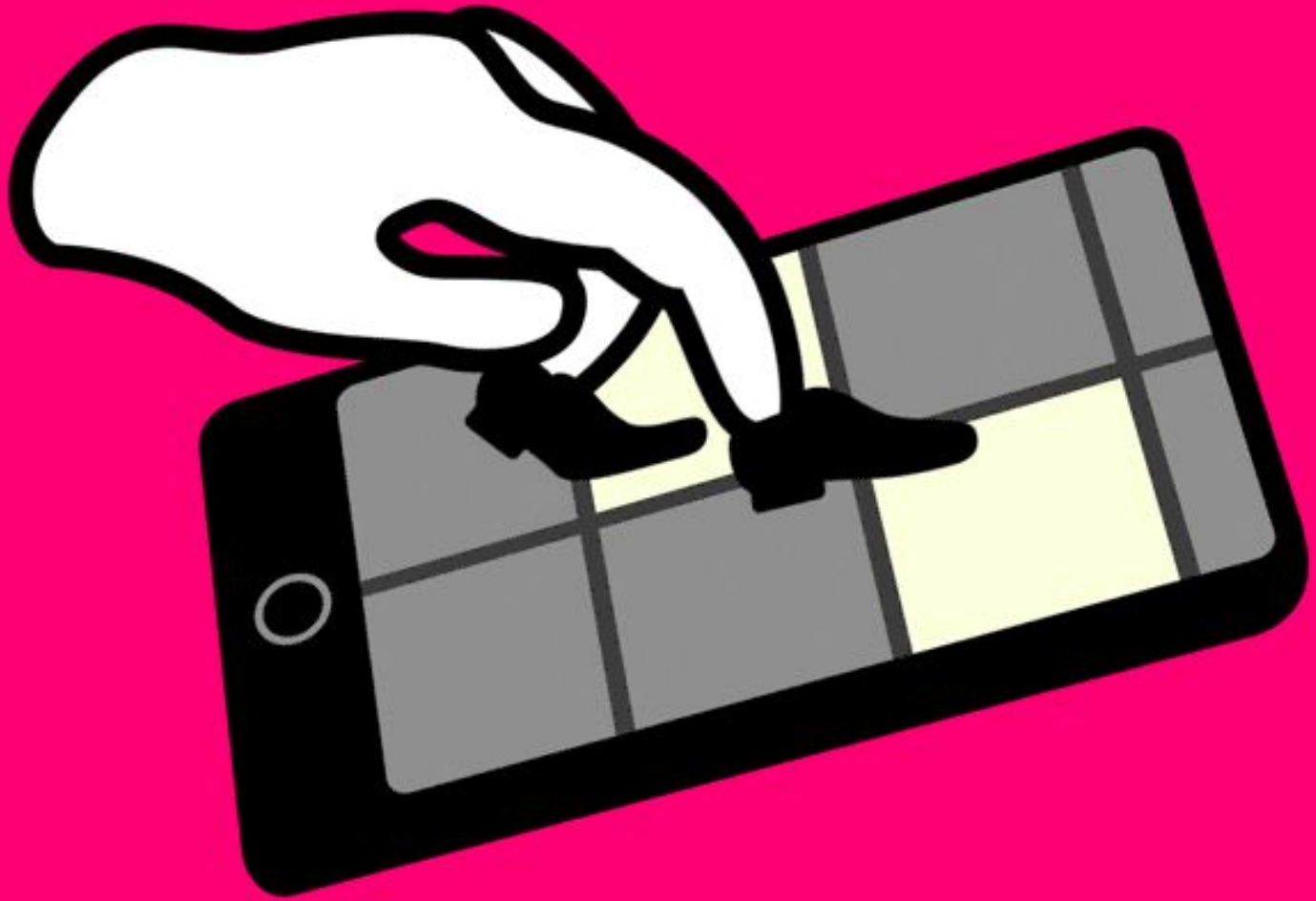


Problem

How do I know what to read?

Problem

News articles are long and boring!

BitNews

BitNews

Bits of News That Matter.

Aditya Bindra
Akshay Bahadur
Naman Arora

Product Canvas

Opportunity

General Demand for efficiency maximization; Saving time amidst information overload is a priority for most

Consumers

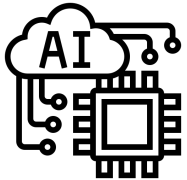
Students and working professionals.
Target consumer: Always short on Time

Strategy

Minimalistic UI
Customizations on-demand
Negligible load times

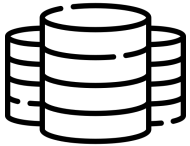
Solution

DistilBERT Model
Extractive Summarizer
Extract the most important information



Data

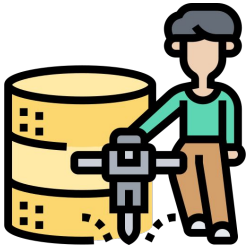
Common Crawl
MIND: Microsoft News Dataset
SQuAD for distillation and testing



Success Criteria

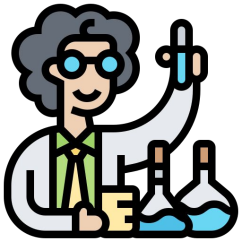
Daily active users and retention
User feedback collection
Time saved by text compression

Team



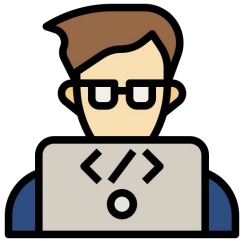
Data Engineering

- Data Warehouse
- Data Pipelines



Data Science

- Model Engineering
- Evaluation
- Metrics



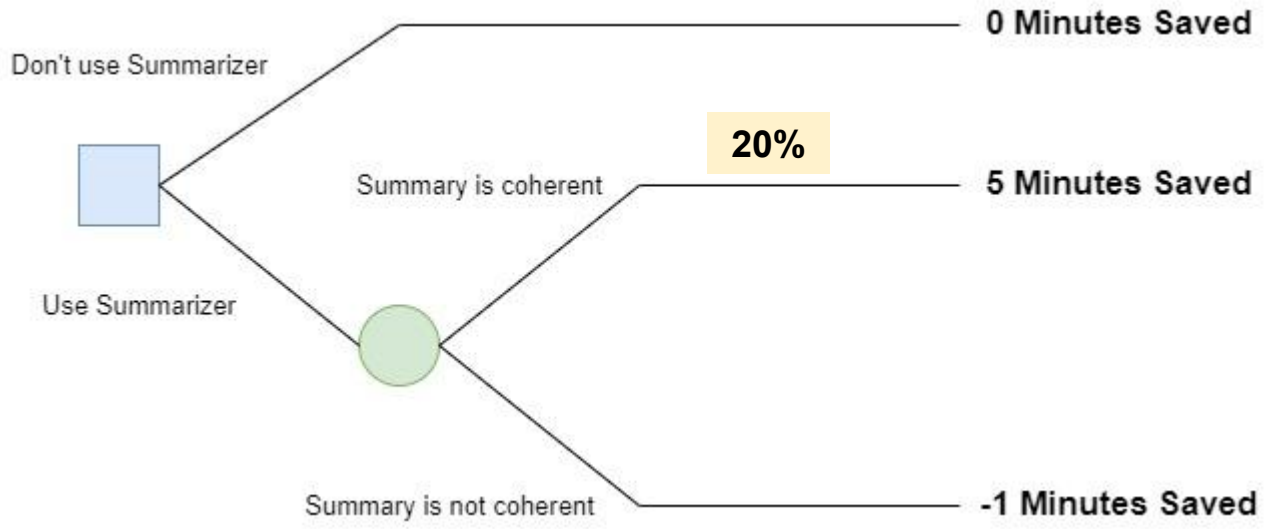
Software Engineering

- Deployment
- Monitoring
- Infrastructure

Product Managers

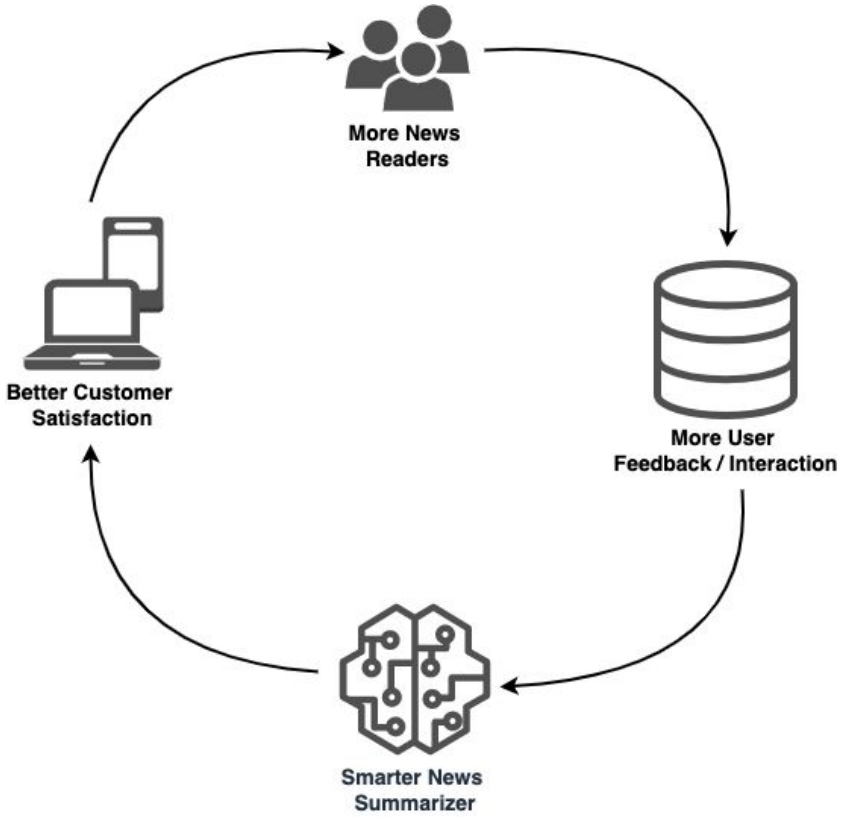
- Business Value Alignment
- Setting goals and vision

Value



Our coherence score: ~85%

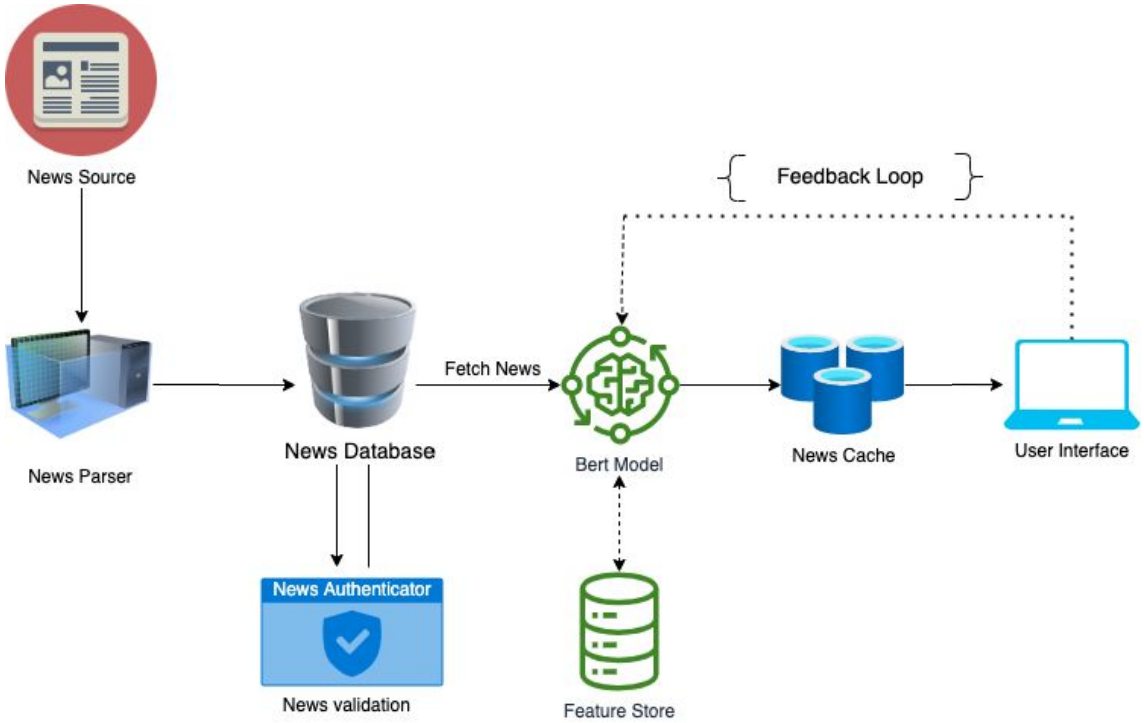
Data Flywheel



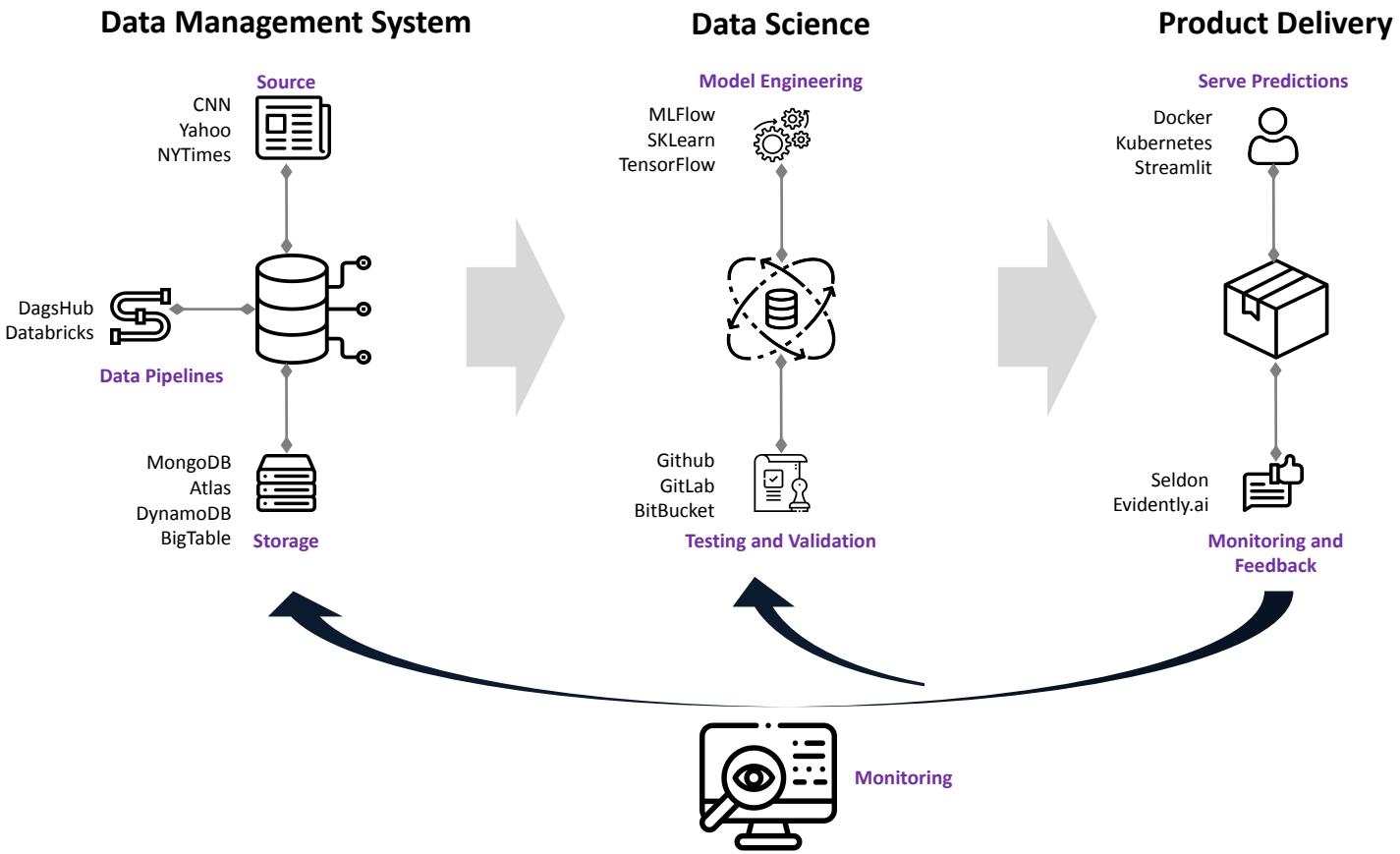
Personalized News Feed

Better Summarization

Architecture



Architecture



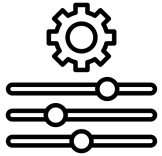
Model

DistilBERT model

- 40% less parameters than BERT
- 60% faster than BERT
- Preserves over 95% of BERT's performance.

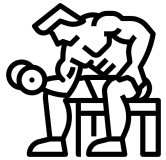
Kullback-Leibler loss (KL Loss)

$$KL(p||q) = \mathbb{E}_p(log(\frac{p}{q})) = \sum_i p_i * log(p_i) - \sum_i p_i * log(q_i)$$



Parameters

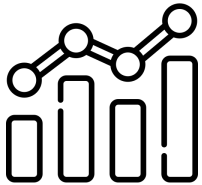
- 66 Billion Parameters
- Average Inference Time = 410 seconds



Bert Extractive Summarizer

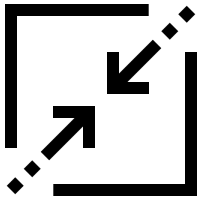
- Pretrained Model for Text Summarization
- Embed the sentences, clustering algorithm to find the sentences that are closest to the cluster's centroids.
- Use Coreference techniques (by HuggingFace) to resolve words in summaries that need more context.

Metrics



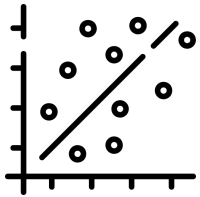
- Dataset : SQuAD 1.1
- F1 score : 85.1
- EM (Exact-match) score : 76.5

Accuracy Measure



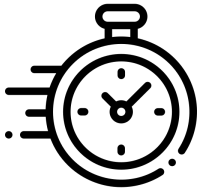
- Metric : Text Compression
- 85-90% text compression

Text-Summary Compression



- Metric : Semantic Cosine Similarity
- 90 - 95% semantic similarity match

Similarity b/w
News and Summary

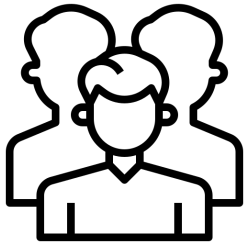


- Metric : Time saved in Minutes
- Median Time saved : 8 minutes
- Maximum Time saved : 57 minutes

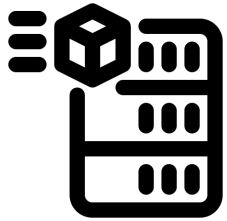
Time Saved per Article

MVP Lessons

Satisfying Information need is Difficult



Every User is Different



Lazy Propagation can be the best option

BitNews

MVP Demo